·数字中医药·

本文引用:周 展, 刘 彬, 郑立瑞, 谭建聪, 邹北骥, 彭清华, 肖晓霞. 一种面向不平衡数据的心脏病风险预测可解释性框架[J]. 湖南中医药大学学报, 2023, 43(6): 1078-1085.

一种面向不平衡数据的心脏病风险预测可解释性框架

周 展[1],刘 彬[1],郑立瑞[1],谭建聪[1],邹北骥[1,2],彭清华[3]*,肖晓霞[1]*
1.湖南中医药大学信息科学与工程学院,湖南 长沙 410208;2.中南大学计算机学院,湖南 长沙 410083;
3.湖南中医药大学中医学院,湖南 长沙 410208

〔摘要〕**目的** 研究疾病预测模型存在的类别不平衡性与不可解释性难题。**方法** 结合极限梯度提升(eXtreme gradient boosting, XGBoost)、混合采样和 Shapley 加法解释(shapley additive exPlanations, SHAP)分析,提出一种面向不平衡数据的心脏病风险预测可解释性框架 ICRPI。**结果** 该框架下的风险预测模型平衡准确度为 0.942 50,AUC 为 0.986 03,模型可视化分析获得高龄、高体质量指数(body mass index, BMI)值、患有糖尿病等 9 个心脏病危险因素,并得出高龄的糖尿病患者、高 BMI 值且诊断为糖尿病或临界糖尿病患者、高 BMI 值且缺乏体力活动群体为患心脏病高危群体,临界糖尿病人群参与体力活动可降低患心脏病风险。**结论** ICRPI 框架适用于真实临床不平衡数据分析,且能明确给出致病风险因素及其相关性,可有效提高临床诊断准确率的同时为医生提供致病因素分析,智能辅助医生临床诊疗。

〔关键词〕数据类别不平衡;心脏病风险预测;XGBoost;SHAP;可解释性
〔中图分类号〕R2 〔文献标志码〕A 〔文章编号〕doi:10.3969/j.issn.1674-070X.2023.06.019

A framework for predicting heart disease risk factors with interpretability by imbalanced data

ZHOU Zhan[1], LIU Bin[1], ZHENG Lirui[1], TAN Jiancong[1], ZOU Beiji[1,2], PENG Qinghua[3], XIAO Xiaoxia[1]**
1. School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China;
2. School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China;
3. School of Chinese Medicine, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

〔**Abstract**〕**Objective** To solve the problems caused by imbalanced data and interpretability of disease prediction models. **Methods** ICRPI, the framework for predicting heart disease risk factors with interpretability by imbalanced data was proposed by combining eXtreme Gradient Boosting (XGBoost), mixed sampling, and Shapley Additive exPlanations (SHAP). **Results** The balance accuracy of the risk prediction model within this framework was 0.942 50, and the AUC was 0.986 03. Nine heart disease factors such as older age, high body mass index (BMI) value, and diabetes were obtained by model visualization analysis. The older diabetic patients, the diabetes or borderline diabetes with high BMI value, the patients with high BMI and lacking physical activities are high-risk groups for heart disease; while for the borderline diabetes, physical activity can reduce the risk of heart disease. **Conclusion** The ICRPI framework can analyze real clinical imbalance data, and can clearly show the pathogenic factors and their correlations. It can effectively improve the accuracy of clinical diagnosis, provide pathogenic factor analysis for doctors, and intelligently assist doctors in clinical practice.

〔**Keywords**〕imbalanced data; predicting heart disease risk factors; XGBoost; SHAP; interpretability

〔收稿日期〕2022-12-03
〔基金项目〕科技部十三五重点研发计划项目(2017YFC1703300);科技创新 2030“新一代人工智能”重大项目课题(2018AAA0102102);2022 年湖南中医药大学研究生创新课题立项基金项目(2022CX123)。
〔第一作者〕周 展,女,硕士研究生,研究方向:类别不平衡下疾病预测模型与模型可解释性。
〔通信作者〕* 肖晓霞,女,博士,副教授,硕士研究生导师,E-mail:amily_x@hnucm.edu.cn;彭清华,男,博士,教授,主治医师,博士研究生导师,E-mail:pqh410007@126.com。

心血管疾病(cardiovascular disease, CVD)是心脏病和血管疾病的一个类别,包括冠心病、脑血管病、先天性心脏病、心力衰竭等。根据《中国心血管健康与疾病报告 2021》推算我国现心血管患病人数为 3.3 亿,2019 年农村和城市 CVD 死亡人数分别占总死亡人数的 46.74%和 44.26%,且死亡率仍处于持续上升趋势[1]。目前,CVD 临床诊断多采用临床血管造影术和影像诊断,该方式对医院资源配置要求较高,检查费用昂贵且对人体有一定创伤[2]。CVD 治疗费用高昂,2019 年中国心脑血管疾病患者的住院总费用为 3 133.66 亿元,且其负担持续加重,特别是在农村地区[1]。因此,早预防、早发现和早治疗是减轻患者负担的关键。

CVD 往往是多种危险因素协同作用的结果,通过疾病风险评估可了解患病风险,做到早预防和早治疗,但这要求医生具有较高专业水平。基于大数据建立机器学习模型并分析患病危险因素可辅助医生诊断决策,提高诊断准确率,缓解医疗资源不均衡问题。同时,也可从海量临床数据中发现疾病诊疗新知识,丰富临床诊断知识。朱宵彤等[2]提出了基于一维卷积的 CVD 预测模型,在尔湾加州大学两个数据集上的独立实验准确率分别 93.36%和 94.48%。李瑞等[3]基于心脑一体化 CT 血管成像预测主要心血管不良事件,采用多因素逻辑回归(logistic regression, LR)分析其危险因素,显示多因素综合的心脑血管系统的影像评估模型预测结果最佳。然而这些研究都是基于类别平衡的数据集,而真实临床数据多为类别不均衡数据,基于这种数据构建的机器学习模型性能较差,且大多机器学习和深度学习模型都缺乏可解释性,无法直接给出模型基于哪些因素进行预测,这将无法满足医疗领域要求模型可解释的需求。目前,CVD 风险预测模型对类别不均衡和模型可解释性的问题关注较少,本文提出基于类别不平衡数据集的 ICRPI 心血管疾病风险预测模型,该模型融合 SMOTEENN 采样、极限梯度提升(eXtreme gradient boosting, XGBoost)等模型和 SHAP 可解释性分析,可获得较高风险预测准确率的同时获得影响 CVD 的危险因素,为构建智能诊疗模型打下基础。

1 相关工作

1.1　基于结构化数据的分类算法

LR 的本质是对数几率(log odds)的线性模型,由于线性模型由特征权重的线性加权组成,可通过特征权重来解释特征对输出的贡献程度,LR 则可通过特征改变带来的对数几率的变化来解释模型,因此 LR 具有内置可解释性。MCRAE 等[4]通过 LR 的可解释性,建立基于多变量指数测定系统的“心脏病计分卡”,分析 CVD 的危险因素,疾病预测心脏健康和心力衰竭的 AUC 分别为 0.840 3 和 0.941 2。决策树(decision tree, DT) 是通过树形结构形象地模拟出决策过程,从根节点到叶子节点的路径代表一条决策路径,只要将 DT 可视化即可了解模型决策全过程,因此 DT 是内置可解释性模型。但为提高模型准确性所建立的DT,往往因层数较深使人类无法真正理解。BLANCO-JUSTICIA 等[5]通过微聚合结合浅层 DT 进行机器学习模型解释。这些具有内置可解释性的分类算法,虽具有较好的模型可解释性,但受模型本身的限制使其预测准确性不高。随机森林(random forest, RF)是由 BREIMAN 等[6]提出的基于 Bagging 的集成学习方法,而 XGBoost 是基于 Boosting 的集成学习算法[7]。相较于其他机器学习模型,不少学者发现集成学习模型预测能力更强[8-11]。但集成学习模型作为“黑盒模型”,在可解释方面存在不足。深度学习模型在图像和自然语言处理领域不仅预测性能高,还能避免大量特征工程工作,但在结构化数据任务中的表现却不如集成树模型。为提升神经网络模型在结构化数据中的性能,许多学者研究如何实现模拟树结构的神经网络架构[12-13]。TabNet 是 Google 发布的针对结构化数据的神经网络模型[14]。与之前学者研究的模型相比,该模型不仅预测性能更好,且可提供模型输出的可视化解释。刘玉航[15]在研究中医哮喘辨证分型中提出基于定向正则化的 TabNet 模型,其辨证模型在多评价指标中占优。尽管基于结构化数据的深度学习模型在近年来取得了较大的进展,但这些模型不论在准确性、性能还是可解释性方面仍然有待改进。从整体上来看,在结构化数据领域分类模型中集成树模型仍然处于优势地位[16]。

1.2　SHAP 可解释性分析

尽管集成树模型在预测能力上取得巨大成功,但缺乏可解释性的模型仍难以在业界应用,尤其在医疗领域中。集成树模型的解释方法常用有两种,第一种方法是将模型转化为可解释的模型,用可解释性模型替代“黑盒模型”进行模型解释。SAGI 等[17]通过将任意决策森林模型转为可解释性 DT 进行模型解释,使其预测能力近似 XGBoost 模型且具有可解释性。这种方法虽然可解释模型,但在预测能力方面仍不足原生模型。第二种方法是使用模型无关的方

法,该方法通过关注模型的输入和输出行为而不是模型的内部结构来解释模型。传统的模型无关解释方法是输出置换特征重要性,这种方法能输出对模型影响较大的特征并直观地反映特征的重要程度,但无法提供具体特征与预测输出的关系,在可解释力度上仍有不足。该问题的替代方案是使用 Shapley 值替代置换特征值,Shapley 值不仅能表示特征重要度还能显示特征如何影响模型输出,例如在二分类任务中,通过输出某特征的 Shapley 值即可表示该特征对模型输出结果值(正类或负类)的贡献度。与传统的特征重要性方法相比,Shapley 值更具有数学上的有效性,它是唯一满足效益性、对称性、虚拟性和可加性的归因方法[18]。但计算 Shapley 值的时间复杂度较高,使其难以应用于真实领域。SHAP 是 Shapley 值的另一种估计方法,该方法极大地提升了计算速度,实现了工业化应用[19]。为计算特征 x的 SHAP 值,假设 set 代表特征 x 与其他特征的所有可能的组合,F 代表所有特征的个数,模型在包含特征 x 的特征组合下的预测结果表示为$Predict_{set}$ (x),模型在不包含特征 x 的特征组合下的预测结果表示为 $Predict_{set/feature}$(x),特征 x 的 SHAP 值计算公式如(1)所示。

$$SHAP_{feature}(x)=\sum_{set:feature\in set}[\,|set|\times\binom{F}{|set|}]^{-1}$$
$$[predict_{set}(x)-predict_{set/feature}(x)] \quad (1)$$

SHAP 概要图是将输出重要特征和特征效应相结合的全局可解释性方法,通过 SHAP 概要图可直观了解每个重要特征对模型类别输出的影响程度,但它无法展示不同特征值下模型输出结果的变化趋势。SHAP 依赖图可展示单个特征取不同值时 SHAP 值的变化趋势,也是一种全局可解释性方法。这两种全局解释方法中,概要图显示重要特征对模型输出的关系,依赖图则从某个重要特征入手进一步展示该特征不同取值时对模型预测的影响。SHAP 可解释性分析属于模型无关可解释方法,相较于传统方法的优势在于具有灵活性且不影响模型的预测能力。

1.3 不平衡分类的数据采样方法

在医疗领域,由于疾病的发病率不同,使得收集到的数据往往存在类别不平衡的问题。传统机器学习方法和深度学习方法在数据类别均衡时能取得较好成绩,相反往往性能极差,特别在类别极度不均衡时。面对类别极度不平衡的数据,通常使用数据采样的方法来解决。主流的采样方法有欠采样、过采样和混合采样,目的都是通过改变数据量使不同类别的样本量达到平衡。欠采样是减少多数类的样本量确保样本量均衡的方法,随机欠采样通过随机丢弃部分多数类样本使样本量达到平衡,是经典的欠采样方法。过采样与欠采样相反,是通过数学模型或方法合成的方式增加少数类样本量使不同类别的样本量均衡。最为经典的过采样方法是 CHAWLA 等[20]提出的 SMOTE 算法,该方法增加了数据量使数据达到均衡,同时提高了数据质量,在诸多领域得到认可。由于过采样能增加样本量则更多应用于小样本数据集,但样本合成的方式容易造成过拟合。混合采样是将欠采样和过采样相结合使不同类别样本量达到平衡的方法,BATISTA 等[21]提出的 SMOTETomek 和 SMOTEENN 算法是较为经典的混合采样方法。混合采样可弥补欠采样导致的样本量减少, 同时能优化过采样导致的样本重叠问题,能在不改变数据量的条件下均衡数据集。

2 对象和方法

2.1 研究对象及数据规范

本文采用 kaggle 网站 2020 年的 Personal Key Indicators of Heart Disease 数据集(https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease),其数据总量为 319 795,包括 167 805 名女性和 151 990 名男性,分类标签为是否患有心脏病,包括 27 373 名心脏病患者和 292 422 名非心脏病患者,共有 17 个特征。通过样本量分析,该数据集具有数据量大且数据类别不平衡的特点。

本文采用的数据集的数据规范化包括:分类标签数值化处理(心脏病患者标记为“1”,非心脏病患者标记为“0”)、特征数值化处理(文本特征值数值化)以及范围特征取均值(如:年龄范围为 55~59,则取 57),规范化结果如表 1 所示。

2.2 ICRPI 框架

本文首先进行数据规范化,将规范后数据集进行类别平衡处理,得到多个“人工”数据集,对各“人工”数据集分别构建模型并得到疾病预测结果,选择最优预测模型并根据心血管医学理论为基准进行 SHAP 分析,具体模型架构如图 1 所示。

ICRPI 框架执行步骤:(1)使用规范化后数据集 S(样本量为 n),由特征集与分类标签组成;(2)对数据集 S 分别进行随机欠采样、SMOTE 过采样、SMOTETomek 和 SMOTEENN 混合采样,得到采样后“人工”数据集;(3)对“人工”数据集分别构建 LR、RF、XGBoost、TabNet 模型,并采用平衡准确度、精度、

表 1　特征规范化

特征	描述	特征值
体质量指数(BMI)	体质量指数数值	[12.02,…,94.85]
吸烟(Smoking)	吸烟	1
	不吸烟	0
饮酒(Alcohol Drinking)	饮酒	1
	不饮酒	0
中风(Stroke)	患有中风	1
	未患中风	0
健康状况(Physical Health)	过去 30 d 中有多少天健康状况不好	[0,1,…,30]
心理健康(Mental Health)	过去 30 d 中有多少天心理健康不好	[0,1,…,30]
行走困难(Diff Walking)	走路或爬楼梯有严重困难	1
	走路或爬楼梯没有严重困难	0
性别(Sex)	男	1
	女	0
年龄范围(Age Category)	年龄区间范围	[21,…,80]
种族(Race)	美国印第安人	0
	亚洲人	1
	黑人	2
	西班牙人	3
	其他	4
	白人	5
糖尿病(Diabetic)	未患糖尿病	0
	妊娠期糖尿病	1
	临界糖尿病	2
	糖尿病	3
体力活动(Physical Activity)	过去 30 d 内除正常工作之外参与身体活动或锻炼	1
	过去 30 d 内除正常工作之外未参与身体活动或锻炼	0
总体健康水平(Gen Health)	出色	0
	非常好	1
	好	2
	一般	3
	不好	4
睡眠时长(Sleep Time)	平均睡眠时长/h	[0,…,23]
哮喘(Asthma)	患有哮喘	1
	未患哮喘	0
肾脏疾病(Kidney Disease)	患有肾脏疾病	1
	未患肾脏疾病	0
皮肤癌(Skin Cancer)	患有皮肤癌	1
	未患皮肤癌	0

召回率、F1 和 AUC 进行模型结果评价，对比模型评价结果，得到预测性能最佳模型 M；(4)通过 SHAP 获取影响模型 M 输出的重要特征；(5)使用 SHAP概要图导出重要特征与心脏病患病的相关关系；(6)使用 SHAP 依赖图导出 top5 重要特征的单变量依赖图；(7)对单变量依赖图进行分析，并导出无法直接反映线性关系的重要特征的交互依赖图；(8)结合相关临床研究结果与实际情况对步骤(6)和步骤(7)中导出的图进行可解释性分析。

3　实验结果与可解释性分析

3.1　实验结果分析与对比

本文以 TabNet、RF、DT、LR 和 XGBoost 模型为基础构建心脏病风险预测模型，采用平衡准确度、AUC、F1、精度和召回率作为模型评价指标，分别对类别不平衡的原始数据集和采样后的数据集构建模型，结果如表 2 所示。从表 2 可知，所有在原始数据集上构建的模型效果都不好。经分别使用随机采样、

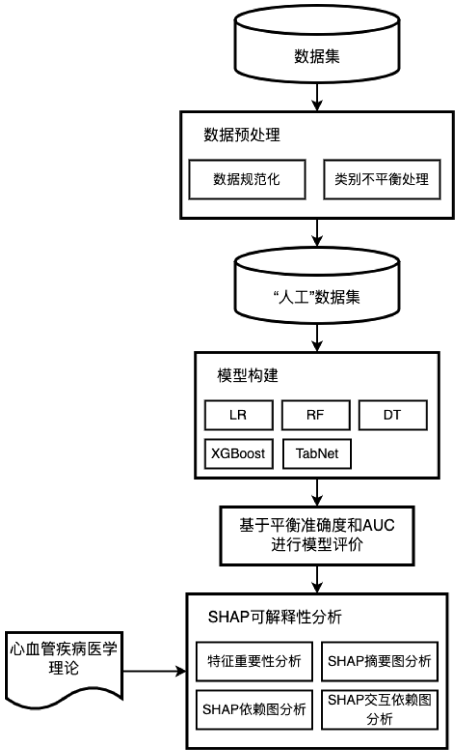

图 1 模型架构图

SMOTE 采样、SMOTETomek 采样和 SMOTEENN 采样后，所建立的模型具有更好的学习能力，其中 XGBoost+SMOTEENN 模型的效果最好，其平衡准确度为 0.942 5，比 RF+SMOTEENN 模型稍好，比 XGBoos 高出 0.41 还多，比 TabNet 模型高出 0.175，说明 XGBoost+SMOTEENN 模型在数据集上效果最佳。

3.2 可解释性分析

由于在采样后建立的 XGBoost+SMOTEENN 模型的预测结果整体要优于其他机器学习模型，因此本文选择该模型做临床诊断可解释性分析。XGBoost 是“黑盒模型”，无法通过模型内置性质获得模型解释，但模型都依赖特征进行预测，可通过分析特征取值与模型预测结果的关系了解模型决策的依据。不同特征对于模型决策的重要程度不同，模型决策结果往往仅受少数重要特征影响，因此，本研究重点分析对模型输出结果影响大的少数重要特征，而不是均摊地解释所有特征。本研究通过 SHAP 值分析训练模型的重要特征信息，模型中特征的 SHAP 值降序排序结果如图 2 所示，展示了从 17 个特征中筛选出的排名 top 9 的重要特征，通过对比不同特征的SHAP 值可知，这些重要特征是对模型输出影响较大的特征。

表 2 基于采样后数据的模型分析结果

模型	平衡准确度	AUC	F1	精度	召回率
TabNet	0.766 88	0.843 39	0.477 15	0.456 35	0.500 00
RF	0.539 07	0.804 69	0.564 55	0.648 20	0.549 45
DT	0.592 65	0.592 68	0.589 55	0.586 80	0.592 65
LR	0.541 38	0.845 35	0.554 80	0.747 45	0.541 40
XGBoost	0.539 07	0.840 46	0.551 35	0.742 40	0.539 05
TabNet+随机欠采样	0.764 50	0.837 25	0.763 25	0.770 55	0.764 90
TabNet+SMOTE	0.792 19	0.875 03	0.789 55	0.791 65	0.789 90
TabNet+SMOTETomek	0.790 77	0.871 62	0.792 35	0.794 00	0.792 70
TabNet+SMOTEENN	0.851 56	0.927 21	0.852 70	0.852 15	0.853 30
RF+随机欠采样	0.743 16	0.811 79	0.743 05	0.744 80	0.743 15
RF+SMOTE	0.885 40	0.953 31	0.885 35	0.886 45	0.885 40
RF+SMOTETomek	0.891 94	0.957 56	0.891 90	0.893 00	0.891 95
RF+SMOTEENN	0.938 86	0.985 93	0.940 55	0.942 80	0.938 85
DT+随机欠采样	0.677 74	0.677 97	0.677 40	0.677 75	0.677 75
DT+SMOTE	0.849 48	0.850 51	0.849 35	0.850 05	0.849 50
DT+SMOTETomek	0.855 04	0.855 95	0.854 95	0.855 70	0.855 05
DT+SMOTEENN	0.908 04	0.908 12	0.909 10	0.910 40	0.908 05
LR+随机欠采样	0.750 63	0.822 99	0.750 25	0.751 05	0.750 65
LR+SMOTE	0.743 74	0.821 40	0.743 35	0.744 95	0.743 75
LR+SMOTETomek	0.748 00	0.823 33	0.747 70	0.749 65	0.748 00
LR+SMOTEENN	0.834 30	0.912 48	0.835 90	0.838 40	0.834 25
XGBoost+随机欠采样	0.755 30	0.829 80	0.764 20	0.766 75	0.765 05
XGBoost+SMOTE	0.911 25	0.969 88	0.912 20	0.912 20	0.912 20
XGBoost+SMOTETomek	0.911 04	0.969 63	0.910 70	0.910 70	0.910 70
XGBoost+SMOTEENN	0.942 50	0.986 03	0.942 80	0.943 10	0.942 50

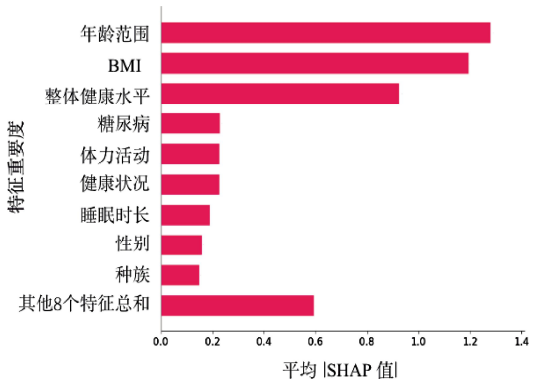

图 2　预测模型特征重要度排名

为进一步明确重要特征对模型输出结果正/负关系，本文使用 SHAP 摘要图进行分析。如图 3所示，SHAP 摘要图显示了模型中的重要特征及对模型预测的影响关系。

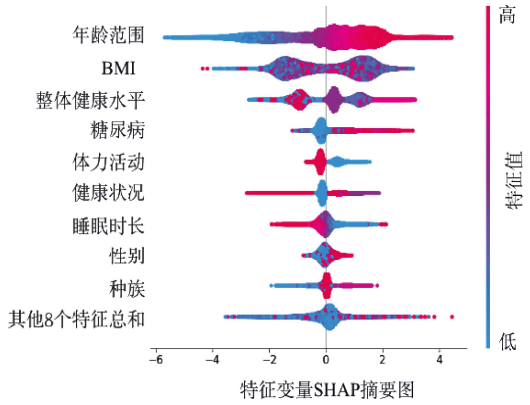

图 3　XGBoost 模型的特征 SHAP 摘要图

注：图中横轴为 SHAP 值，纵轴表示不同特征，从原点向右 SHAP 值为正，表示特征对预测结果的贡献度为正，线条向右越多表示贡献度越大，向左反之，而线条越粗代表样本量越大，反之越小，颜色从蓝到红代表征值取值从小到大。

SHAP 单变量依赖图可分析单个特征与 SHAP 值之间的线性关系，图 4 是对输出结果有影响的排名 top5 的重要特征的单变量依赖图。图 4(a)显示 SHAP 值随年龄增加，中老年人患心脏病风险更高；图 4(b)显示参与体力活动的成年人比缺乏体力活动的成年人 SHAP 值更低，表明缺乏锻炼是导致心脏病的危险因素；图 4(c)显示患有糖尿病和临界糖尿病的成年人有更多患心脏病的风险；图 4(d)表明整体健康水平越高患心脏病风险越低；图 4(e)中“BMI”与 SHAP 值不是简单线性关系，说明该特征可能与其他特征交互影响预测结果，无法通过单变量分析特征与模型的预测关系，需要引入双变量突出组合特征效应的交互依赖图做进一步分析。

本文分析“BMI”分别与“体力活动”“糖尿病”的组合特征关系，如图 5 所示。图 5(a)中显示当“BMI”取值大于 35 时，红色点靠下居多，蓝色点靠上居多，表明“BMI”较高且缺乏体力活动的成年人有更高患病风险。因此，将“BMI”与“糖尿病”组合起来分析，如图 5(b)所示，“BMI”取值大于 30 时，红色点靠右上居多，表明“BMI”较高的糖尿病患者或临界糖尿病患者有更高的患心脏病风险。

从“糖尿病”单变量依赖图可知糖尿病与临界糖尿病患者是患心脏病的高危人群，构建“糖尿病”与“体力活动”及“年龄范围”的交互依赖图进一步分析患病因素，如图 6 所示。图 6(a)显示“糖尿病”取值为 2 时，蓝色区域靠上，说明临界糖尿病且缺乏体力活动的成年人有更高的患心脏病风险。图 6(b)中显示“糖尿病”取值为 3 时，红色区域靠上，表明糖尿病患者随年龄的增长患心脏病风险提高。

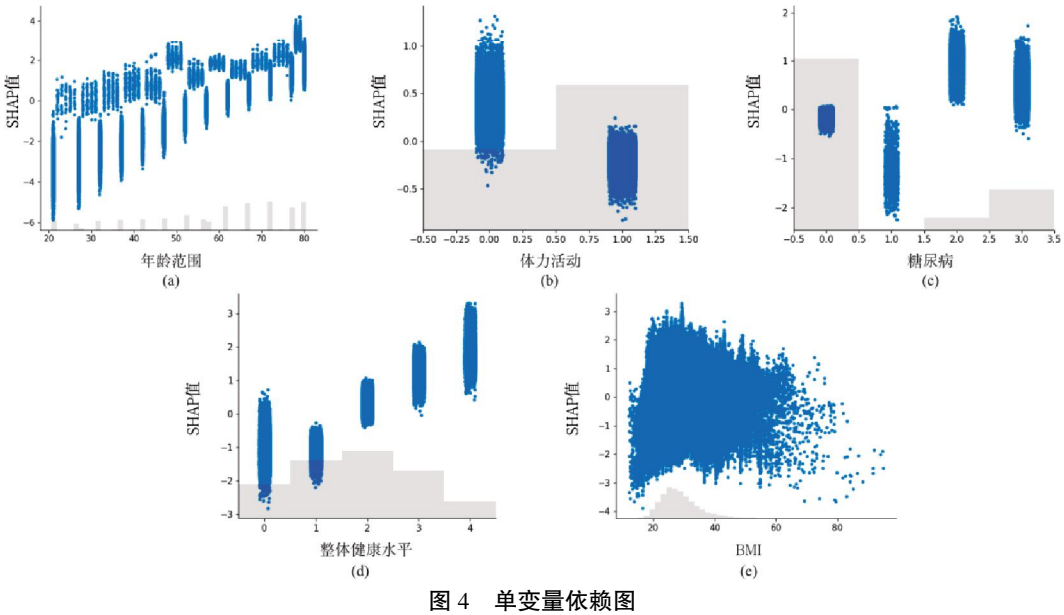

图 4　单变量依赖图

注：纵坐标为 SHAP 值，横坐标为特征取值，单变量依赖图用于描述随特征不同取值时 SHAP 值的变化趋势。

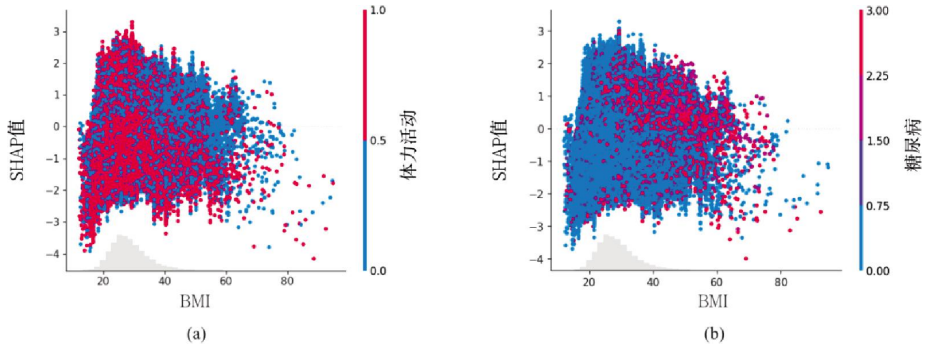

图 5 BMI 交互依赖图

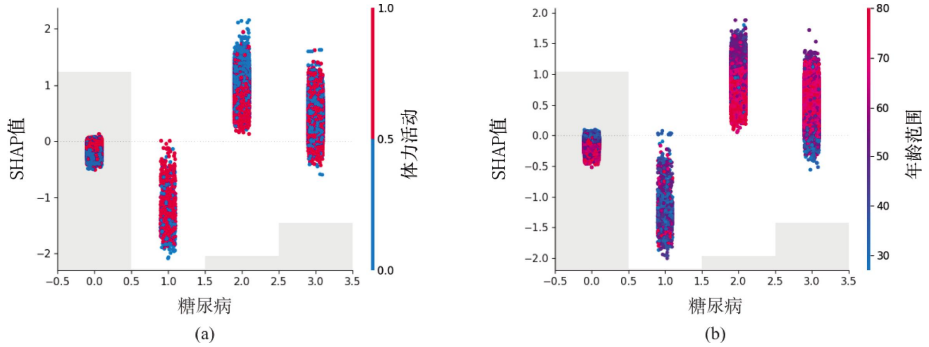

图 6 糖尿病交互依赖图

4 讨论

本文结合 XGBoost 和 SMOTEENN 算法提出了 ICRPI 框架，该框架预测的平衡准确度超过 94%，AUC 值超过 98%，且能提取与患病风险关系紧密的重要特征及其与患病风险的关系。通过本研究的实验得出传统机器学习模型、集成学习和 TabNet 对类别不平衡的数据分类效果不好，混合采样后的集成学习模型分类效果最佳。

为解释模型，本文使用 SHAP 进行可解释性分析。通过 SHAP 值排序获得年龄、高 BMI 值、糖尿病、缺乏体力活动等 9 个模型重要特征，有大量研究表明肥胖、糖尿病、缺乏锻炼等是导致心血管疾病的危险因素[22-24]，这说明通过 SHAP 值筛选出的重要特征符合医学临床真实情况。为获得重要特征与模型输出的关系，本文使用了基于 SHAP 值的单变量依赖图和交互依赖图，通过单变量依赖图得出“年龄范围”“糖尿病”“体力活动”和“整体健康水平”这些特征与心脏病诊断结果存在线性关系。然而，单变量依赖图无法直接得出“BMI”与输出结果的线性关系。为分析“BMI”与其他特征的交互关系，需要了解在医学临床中哪些特征与“BMI”具有相关性。由于肥胖和缺乏锻炼是心脏病的危险因素[22]，而肥胖可体现在较高的“BMI”值上，缺乏锻炼可体现在缺乏一定的体力活动上，于是分析“BMI”与“体力活动”共同作用于心脏病患病风险。ECKEL 等[25]提出 BMI 和糖尿病体现了心脏代谢风险，是引发心脏病的危险指标，因此，将“BMI”与“糖尿病”组合起来分析。从“BMI”交互依赖图的分析得出高 BMI 值是心脏病危险指标，主要体现在高 BMI 值且缺乏体力活动与高BMI 值的糖尿病患者或临界糖尿病人群中，分析结果与临床实际情况相符合。2 型糖尿病防治指南指出，2 型糖尿病患病时长大于等于 10 年或合并年龄大于 50 岁等为心血管风险高危因素，早期生活方式干预(如加强运动)可有效减少 2 型糖尿病的发生或延缓并发症的发展[26]。本文构建“糖尿病”与“体力活动”及“年龄范围”的交互依赖图分析得出高龄糖尿病患者及缺乏锻炼的临界糖尿病人群具有较高心脏病患病风险，该结果与 2 型糖尿病防治指南观点一致。

综上所述，本研究表明 ICRPI 框架可以面向真实的临床类别不平衡数据构建合适的具有较高预测性能的分类模型，且能客观地给出致病因素分析，可辅助医生提高心血管疾病风险预测准确率，降低心脏病诊疗费用并减少人体的创伤。该框架目前仅可解释特征与预测结果的相关性，不能解释特征与预

测结果的因果关系,但这一框架为构建面向真实临床的高准确率、可解释性的风险预测模型提供一种有效途径,满足临床智能诊疗系统需求。

参考文献

[1] 中国心血管健康与疾病报告编写组. 中国心血管健康与疾病报告 2021 概要[J]. 中国循环杂志, 2022, 37(6): 553-578.

[2] 朱宵彤, 庞春颖, 朱　涵. 基于深度学习的心血管疾病预测模型[J]. 计算机应用, 2021, 41(S2): 346-350.

[3] 李　瑞, 刘墨麒, 黎佳璐, 等. 心脑血管系统的影像评估对主要心血管不良事件的预测作用[J]. 中国脑血管病杂志, 2022, 19(3): 154-160.

[4] MCRAE M P, BOZKURT B, BALLANTYNE C M, et al. Cardiac ScoreCard: A diagnostic multivariate index assay system for predicting a spectrum of cardiovascular disease[J]. Expert Systems With Applications, 2016, 54: 136-147.

[5] BLANCO-JUSTICIA A, DOMINGO-FERRER J, MARTíNEZ S, et al. Machine learning explainability via microaggregation and shallow decision trees[J]. Knowledge-Based Systems, 2020, 194: 105532.

[6] BREIMAN L. Random forests[J]. Machine learning, 2001, 45: 5-32.

[7] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York: ACM, 2016: 785-794.

[8] 陈　苗, 陈　青, 尹晓清. 随机森林的集成分类算法对心胸外科 ICU 患者谵妄风险的预测分析[J]. 中国胸心血管外科临床杂志, 2022, 29(7): 886-891.

[9] 郑晓燕. 基于机器学习的心血管疾病预测系统研究[D]. 北京: 北京交通大学, 2018.

[10] 于大海. 基于 BP 神经网络和随机森林算法的冠状动脉狭窄风险识别模型研究[D]. 太原: 山西医科大学, 2019.

[11] 彭佳丽, 刘春容, 李　旭, 等. 采用 XGBoost 和随机森林探索中国西部女性乳腺癌危险因素[J]. 现代预防医学, 2020, 47(1): 1-4.

[12] KONTSCHIEDER P, FITERAU M, CRIMINISI A, et al. Deep neural decision forests[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2016: 1467-1475.

[13] YANG Y, MORILLO I G, HOSPEDALES T M. "Deep neural decision trees"[EB/OL]. 2018. DOI: 10.48550/arXiv.1806.06988.

[14] ARIK S Ö, PFISTER T. TabNet: Attentive interpretable tabular learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(8): 6679-6687.

[15] 刘玉航. 基于机器学习的中医哮喘辨证分型研究与应用[D]. 青岛: 青岛科技大学, 2022.

[16] SHWARTZ-ZIV R, ARMON A. Tabular data: Deep learning is not all you need[J]. Information Fusion, 2022, 81: 84-90.

[17] SAGI O, ROKACH L. Approximating XGBoost with an interpretable decision tree[J]. Information Sciences, 2021, 572: 522-542.

[18] MOLNAR C. Interpretable machine learning:A guide for making black box models explainable[M]. Fletcher, NC, USA: LULU, Feb. 2019: 295-296.

[19] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December, 2017: 4765-4774.

[20] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[21] BATISTA G E A P A, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.

[22] LI X, WU C Q, LU J P, et al. Cardiovascular risk factors in China: A nationwide population-based cohort study [J]. The Lancet Public Health, 2020, 5(12): e672-e681.

[23] 何　源, 马少宁, 王海宏, 等. 宁夏回族自治区心血管疾病高危人群筛查与相关危险因素研究[J]. 现代预防医学, 2022, 49(1): 21-26, 31.

[24] 刘　览, 刘华章, 冯颖青, 等. 广州市 35~75 岁社区居民心血管病主要危险因素聚集情况分析[J]. 现代预防医学, 2020, 47(4): 635-639, 647.

[25] ECKEL R H, KAHN R, ROBERTSON R M, et al. Preventing cardiovascular disease and diabetes: A call to action from the American Diabetes Association and the American Heart Association[J]. Circulation, 2006, 113(25): 2943-2946.

[26] 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南(2020 年版)[J]. 国际内分泌代谢杂志, 2021, 41(5): 482-548.

(本文编辑　匡静之)